In [ ]:
```
# Titanic Dataset - Exploratory Data Analysis (EDA)

## Executive Summary
This project performs Exploratory Data Analysis (EDA) on the Titanic dataset.
Using Python libraries like Pandas, Matplotlib, and Seaborn, we explored passenger demographics, survival distributions,
and relationships between various features. Key insights include higher survival rates among females, passengers from 1st class, and children.
Feature engineering such as family size and extracted titles further revealed patterns influencing survival.
The analysis concludes that socio-economic status, age, and gender played major roles in survival outcomes.
```

In [ ]:
```
## Table of Contents
- [1. Introduction](#introduction)
- [2. Data Overview](#data-overview)
- [3. Missing Values](#missing-values)
- [4. Feature Engineering](#feature-engineering)
- [5. Univariate Analysis](#univariate-analysis)
- [6. Bivariate Analysis](#bivariate-analysis)
- [7. Correlation Heatmap](#correlation-heatmap)
- [8. Advanced Visualizations](#advanced-visualizations)
- [9. Outlier Detection](#outlier-detection)
- [10. Summary Table](#summary-table)
- [11. Observations](#observations)
- [12. Final Summary](#final-summary)
```

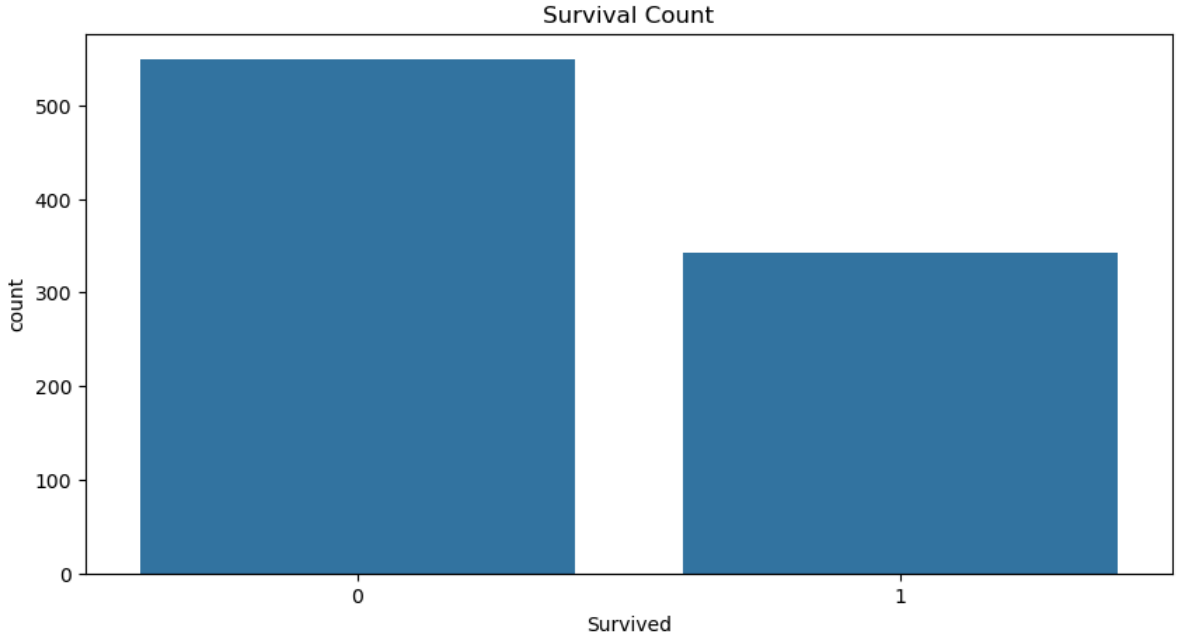
In [5]:
```
# 1. Introduction
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load Dataset
df = pd.read_csv('train.csv')  # Change path if needed

# 2. Data Overview
print("\n--- HEAD ---\n")
print(df.head())

print("\n--- INFO ---\n")
print(df.info())

print("\n--- DESCRIBE ---\n")
print(df.describe())

print("\n--- VALUE COUNTS (Survived) ---\n")
print(df['Survived'].value_counts())

# 3. Missing Values
print("\n--- MISSING VALUES ---\n")
print(df.isnull().sum())

# 4. Feature Engineering
df['FamilySize'] = df['SibSp'] + df['Parch'] + 1
df['IsAlone'] = 1  # Default is alone
df.loc[df['FamilySize'] > 1, 'IsAlone'] = 0  # If family size > 1, not alone
df['Title'] = df['Name'].str.extract(r' ([A-Za-z]+)\.', expand=False)

# Fill missing values
df['Age'] = df['Age'].fillna(df.groupby('Pclass')['Age'].transform('median'))
```

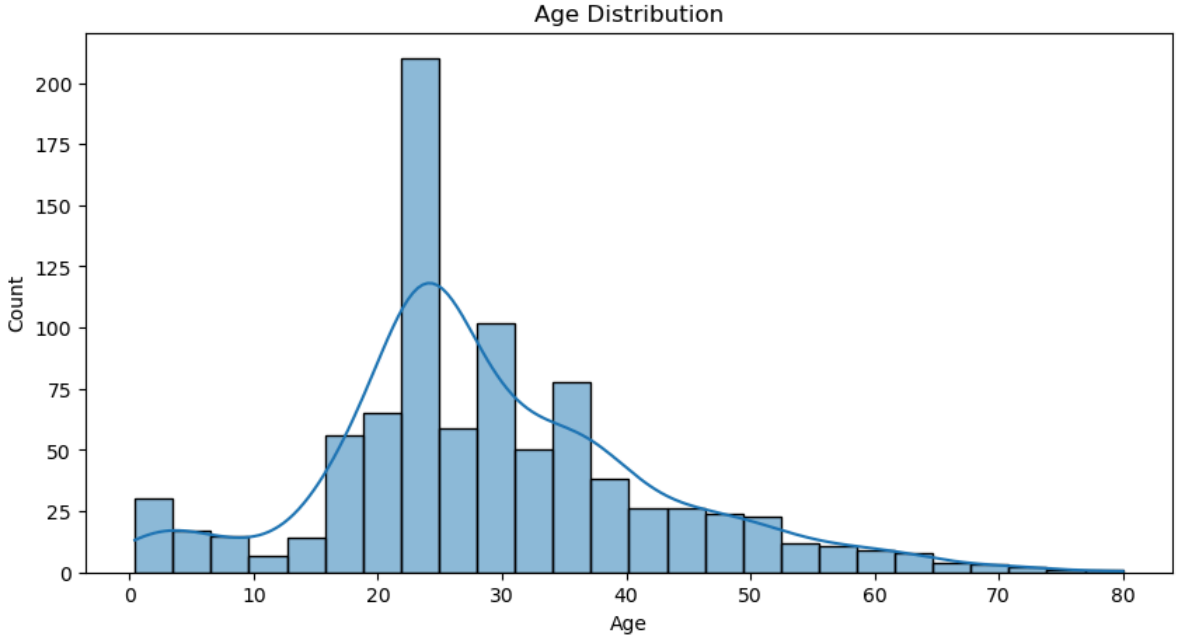
```
df['Embarked'] = df['Embarked'].fillna(df['Embarked'].mode()[0])
df['Fare'] = df['Fare'].fillna(df['Fare'].median())

# 5. Univariate Analysis
plt.figure(figsize=(10,5))
sns.countplot(x='Survived', data=df)
plt.title('Survival Count')
plt.show()

plt.figure(figsize=(10,5))
sns.histplot(df['Age'], kde=True)
plt.title('Age Distribution')
plt.show()

# 6. Bivariate Analysis
plt.figure(figsize=(10,5))
sns.countplot(x='Pclass', hue='Survived', data=df)
plt.title('Passenger Class vs Survival')
plt.show()

plt.figure(figsize=(10,5))
sns.countplot(x='Sex', hue='Survived', data=df)
plt.title('Sex vs Survival')
plt.show()

plt.figure(figsize=(10,5))
sns.countplot(x='Embarked', hue='Survived', data=df)
plt.title('Embarked vs Survival')
plt.show()

plt.figure(figsize=(10,5))
sns.boxplot(x='Survived', y='Fare', data=df)
plt.title('Fare vs Survival')
plt.show()

# 7. Correlation Heatmap (fixed)
plt.figure(figsize=(10,8))
sns.heatmap(df.select_dtypes(include=['int64', 'float64']).corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()

# 8. Advanced Visualizations
plt.figure(figsize=(10,5))
sns.violinplot(x='Pclass', y='Age', hue='Survived', data=df, split=True)
plt.title('Pclass and Age vs Survival')
plt.show()

plt.figure(figsize=(10,5))
sns.scatterplot(x='Age', y='Fare', hue='Survived', data=df)
plt.title('Age vs Fare Scatter Plot')
plt.show()

sns.pairplot(df[['Survived', 'Pclass', 'Age', 'Fare', 'FamilySize']], hue='Survived')
plt.show()

# 9. Outlier Detection
plt.figure(figsize=(10,5))
sns.boxplot(x='Fare', data=df)
```

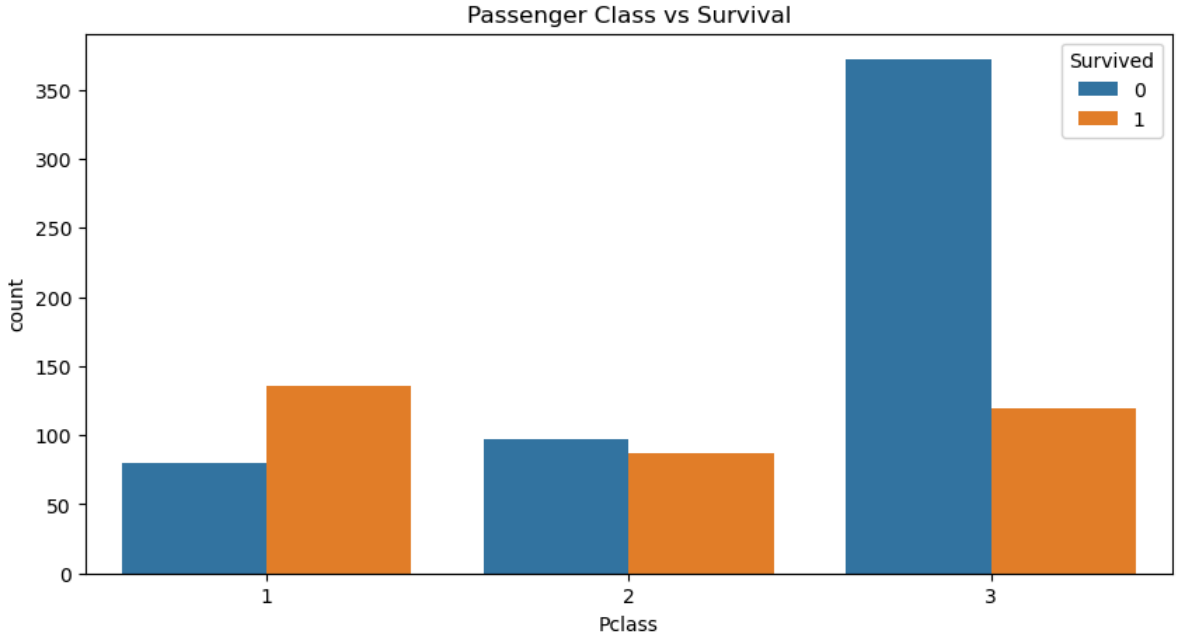
```
plt.title('Fare Outliers')
plt.show()

plt.figure(figsize=(10,5))
sns.boxplot(x='Age', data=df)
plt.title('Age Outliers')
plt.show()

# 10. Summary Table
summary = pd.DataFrame({
    'Feature': ['Sex', 'Pclass', 'Age', 'Fare', 'Embarked', 'FamilySize', 'Title'],
    'Key Insight': [
        'Females survived more',
        '1st class had better survival',
        'Children had slightly better survival',
        'Higher fare passengers survived more',
        'Embarked from C had higher survival',
        'Large families had less survival chances',
        'Certain titles (e.g., Mrs) survived more'
    ]
})

print("\n--- SUMMARY TABLE ---\n")
print(summary)

# 11. Observations
print("\n--- OBSERVATIONS ---\n")
print("- Most passengers did not survive.")
print("- Females had a higher survival rate than males.")
print("- Passengers in 1st class had higher survival rates.")
print("- Younger passengers (especially children) had better chances.")
print("- Fare prices and Embarked port also impacted survival.")
print("- Passengers traveling alone were less likely to survive.")

# 12. Final Summary
print("\n--- FINAL SUMMARY ---\n")
print("Survival was highly influenced by gender, class, age, fare, and family connections. New features like FamilySize and Title improved understanding of survival factors.")
```

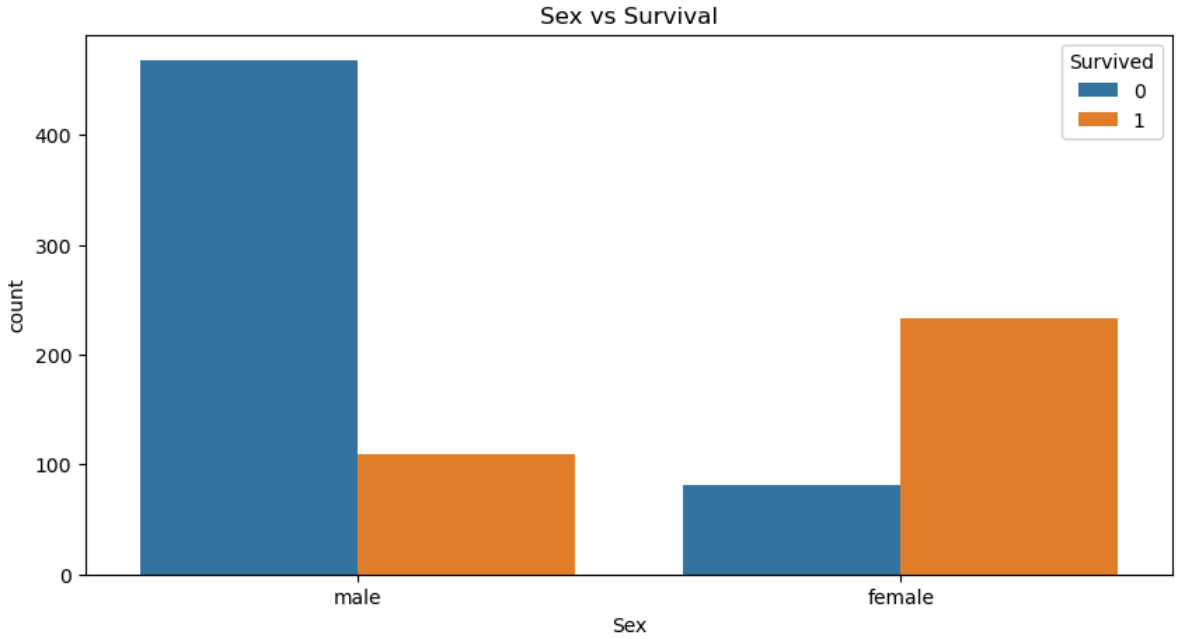
```
--- HEAD ---

   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3

                                                Name     Sex   Age  SibSp  \
0                            Braund, Mr. Owen Harris    male  22.0      1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1
2                             Heikkinen, Miss. Laina  female  26.0      0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                           Allen, Mr. William Henry    male  35.0      0

   Parch            Ticket     Fare Cabin Embarked
0      0         A/5 21171   7.2500   NaN        S
1      0          PC 17599  71.2833   C85        C
2      0  STON/O2. 3101282   7.9250   NaN        S
3      0            113803  53.1000  C123        S
4      0            373450   8.0500   NaN        S

--- INFO ---

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
None

--- DESCRIBE ---

       PassengerId    Survived      Pclass         Age       SibSp  \
count   891.000000  891.000000  891.000000  714.000000  891.000000
mean    446.000000    0.383838    2.308642   29.699118    0.523008
std     257.353842    0.486592    0.836071   14.526497    1.102743
min       1.000000    0.000000    1.000000    0.420000    0.000000
25%     223.500000    0.000000    2.000000   20.125000    0.000000
50%     446.000000    0.000000    3.000000   28.000000    0.000000
75%     668.500000    1.000000    3.000000   38.000000    1.000000
max     891.000000    1.000000    3.000000   80.000000    8.000000
```

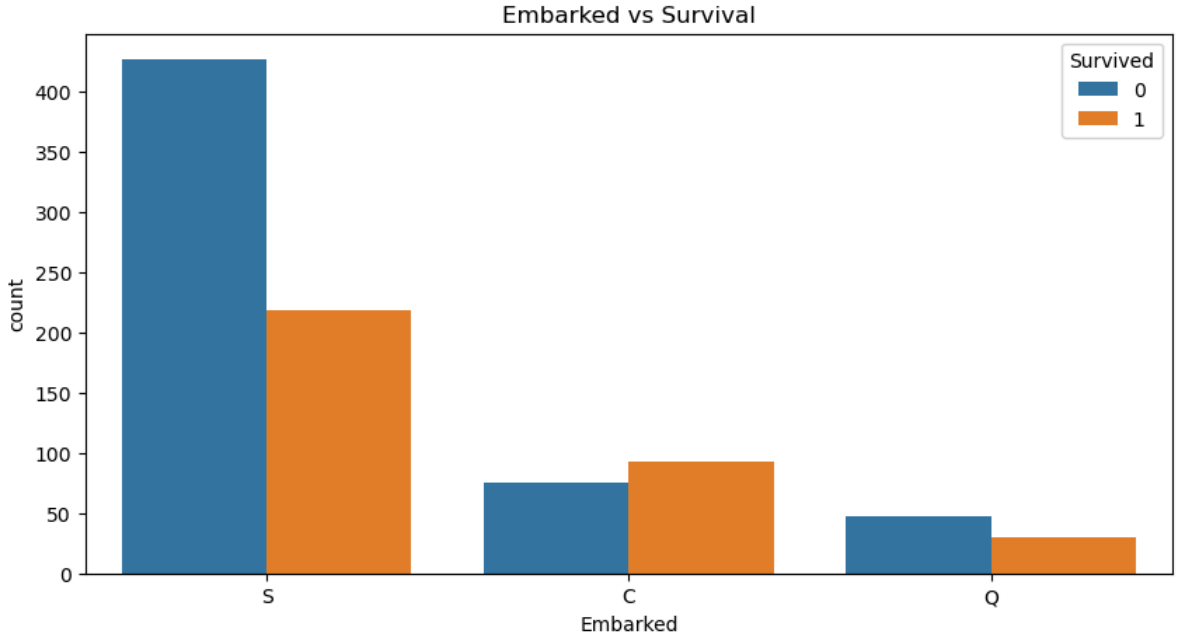
```
            Parch        Fare
count  891.000000  891.000000
mean     0.381594   32.204208
std      0.806057   49.693429
min      0.000000    0.000000
25%      0.000000    7.910400
50%      0.000000   14.454200
75%      0.000000   31.000000
max      6.000000  512.329200

--- VALUE COUNTS (Survived) ---

Survived
0    549
1    342
Name: count, dtype: int64

--- MISSING VALUES ---

PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

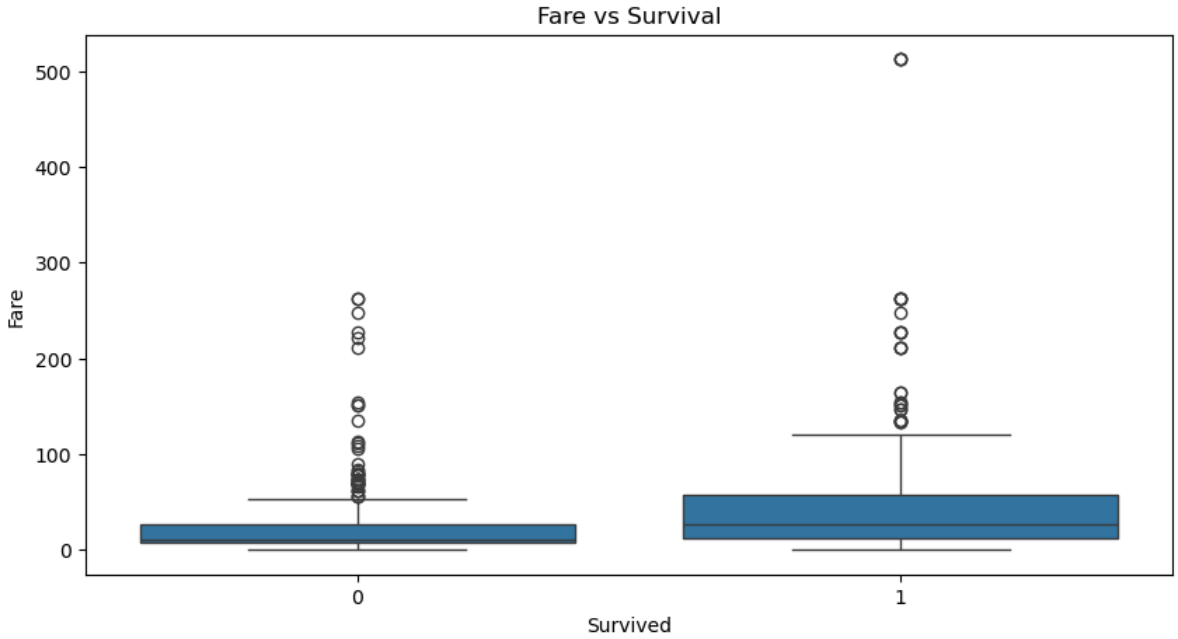

Survival Count
count
500
400
300
200
100
0
0
1
Survived

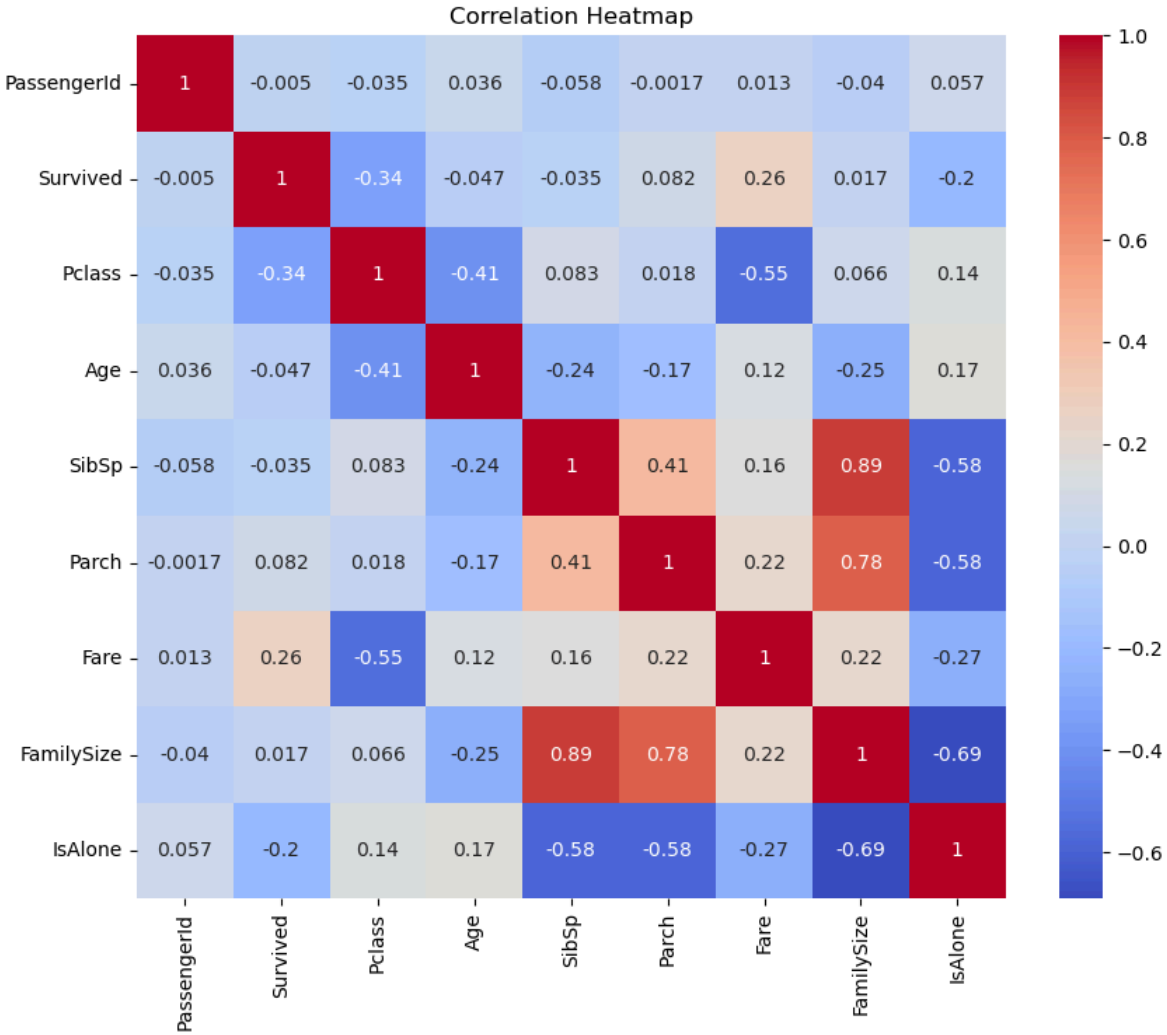

Age Distribution
Count
200
175
150
125
100
75
50
25
0
0
10
20
30
40
50
60
70
80
Age


Passenger Class vs Survival
Survived
0
1
count
0
50
100
150
200
250
300
350
1
2
3
Pclass

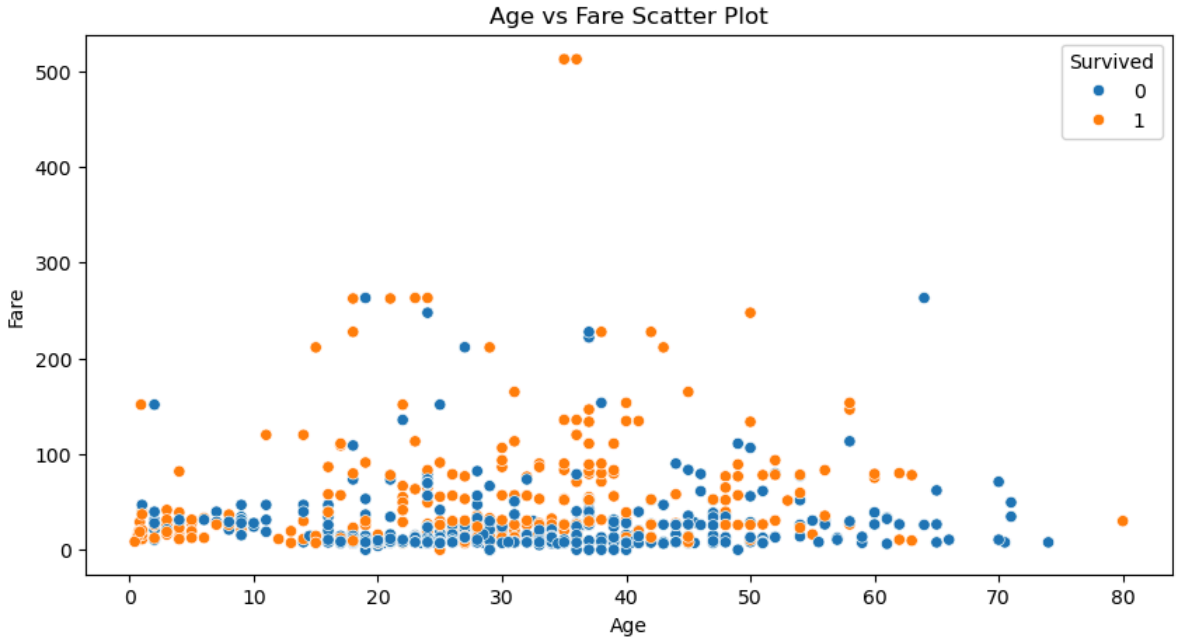

Sex vs Survival
Survived
0
1
count
0
100
200
300
400
male
female
Sex

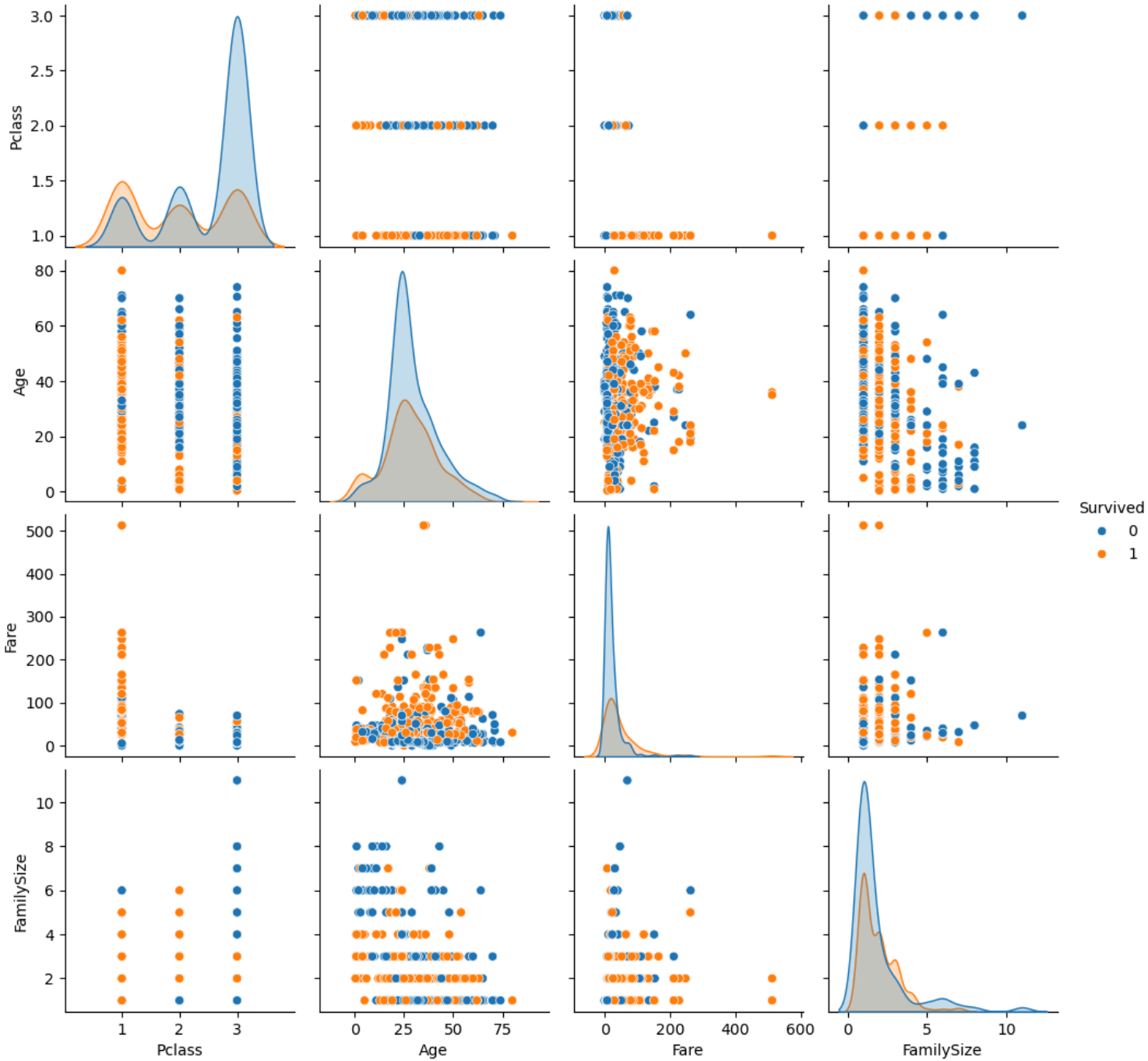

Embarked vs Survival
Survived
0
1
count
400
350
300
250
200
150
100
50
0
S
C
Q
Embarked

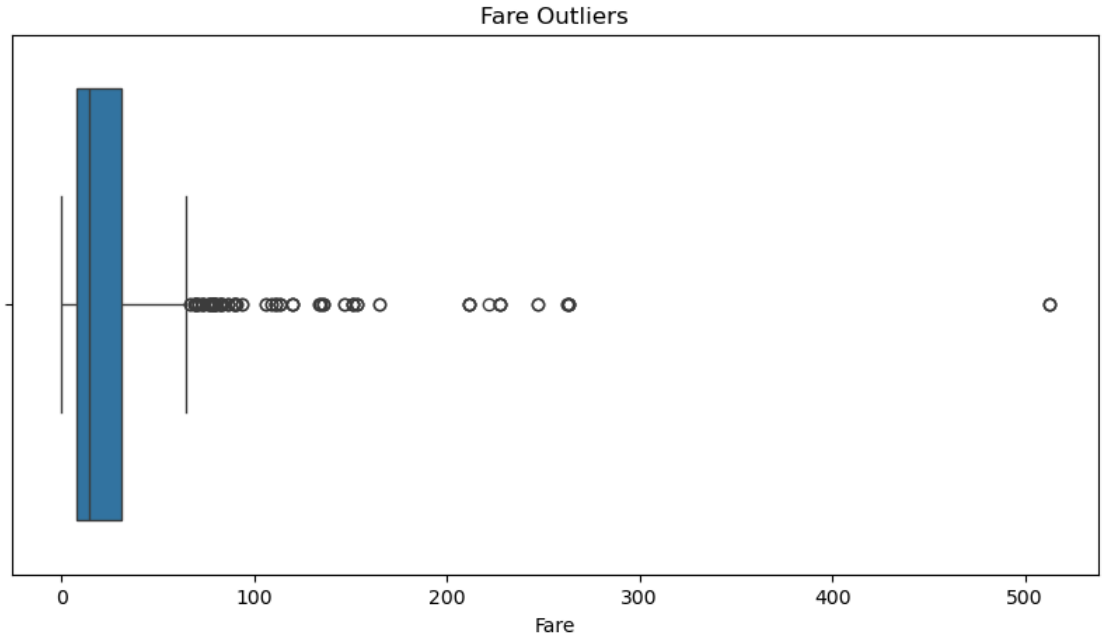

Fare vs Survival
Fare
500
400
300
200
100
0
0
1
Survived

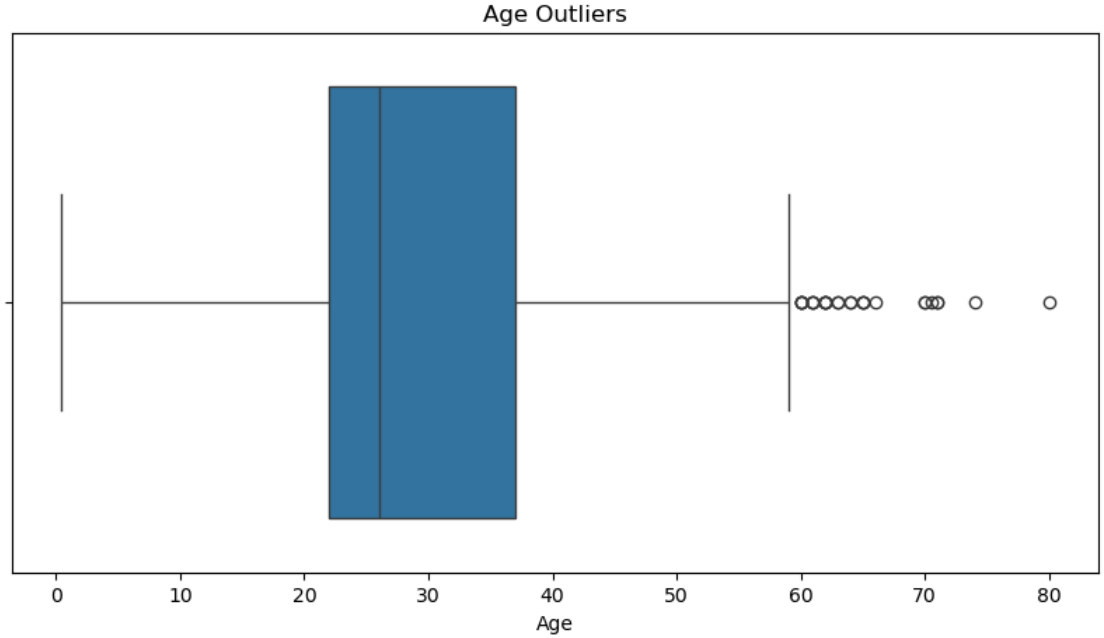

Correlation Heatmap
PassengerId 1 -0.005 -0.035 0.036 -0.058 -0.0017 0.013 -0.04 0.057
Survived -0.005 1 -0.34 -0.047 -0.035 0.082 0.26 0.017 -0.2
Pclass -0.035 -0.34 1 -0.41 0.083 0.018 -0.55 0.066 0.14
Age 0.036 -0.047 -0.41 1 -0.24 -0.17 0.12 -0.25 0.17
SibSp -0.058 -0.035 0.083 -0.24 1 0.41 0.16 0.89 -0.58
Parch -0.0017 0.082 0.018 -0.17 0.41 1 0.22 0.78 -0.58
Fare 0.013 0.26 -0.55 0.12 0.16 0.22 1 0.22 -0.27
FamilySize -0.04 0.017 0.066 -0.25 0.89 0.78 0.22 1 -0.69
IsAlone 0.057 -0.2 0.14 0.17 -0.58 -0.58 -0.27 -0.69 1
PassengerId
Survived
Pclass
Age
SibSp
Parch
Fare
FamilySize
IsAlone
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


Pclass and Age vs Survival
Survived
0
1
Age
Pclass
1
2
3
0
20
40
60
80


Age vs Fare Scatter Plot
Survived
0
1
Fare
Age
0
100
200
300
400
500
0
10
20
30
40
50
60
70
80


Pclass
Age
Fare
FamilySize
Survived
0
1


Fare Outliers
0
100
200
300
400
500
Fare


Age Outliers
0
10
20
30
40
50
60
70
80
Age

```
--- SUMMARY TABLE ---

      Feature                              Key Insight
0         Sex                    Females survived more
1      Pclass            1st class had better survival
2         Age     Children had slightly better survival
3        Fare      Higher fare passengers survived more
4    Embarked       Embarked from C had higher survival
5  FamilySize  Large families had less survival chances
6       Title  Certain titles (e.g., Mrs) survived more

--- OBSERVATIONS ---

- Most passengers did not survive.
- Females had a higher survival rate than males.
- Passengers in 1st class had higher survival rates.
- Younger passengers (especially children) had better chances.
- Fare prices and Embarked port also impacted survival.
- Passengers traveling alone were less likely to survive.

--- FINAL SUMMARY ---

Survival was highly influenced by gender, class, age, fare, and family connections. New features like FamilySize and Title improved understanding of survival factors.
```

In [ ]: